Efficiently estimating some common geostatistical models from "image-type, possibly incomplete" datasets: CGEMEV and its extension to unknown nugget-effect

Didier A. Girard, Rémy Drouilhet

▶ **To cite this version:**

Didier A. Girard, Rémy Drouilhet. Efficiently estimating some common geostatistical models from "image-type, possibly incomplete" datasets: CGEMEV and its extension to unknown nugget-effect. Spatial Statistics 2019 : Towards Spatial Data Science, Jul 2019, Sitgès, Barcelone, Spain. hal-02174478v2

HAL Id: hal-02174478
https://hal.science/hal-02174478v2

Submitted on 10 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficiently estimating some common geostatistical models from «image-type, possibly incomplete» datasets: CGEMEV and its extension to unknown nugget-effect

Rémy Drouilhet[1], Didier Girard[1,2]

[1]Univ. Grenoble Alpes, LJK, France, [2]CNRS, LJK, France

The problem of estimating the 2 parameters ("**range**" and "**variance**") of a stationary isotropic Gaussian process whose autocorrelation function belongs to the Matérn class, appears in many contexts (e.g. [9,11]). We propose two extensions of the CGEM-EV method (cf. [5,6]) for the important case of **unknown "nugget-variance"** (3rd parameter), and compare them to the classic maximum likelihood (ML) method, for the Matérn subclass $\nu = 1/2$, and "small image"-type observations

Background

We mainly consider the following statistical model which arises e.g. in remote sensing image analysis: let $Z(\mathbf{s})$, $\mathbf{s} \in \mathbb{R}^2$, be a zero mean stationary Gaussian stochastic process whose autocorrelation function is assumed to belong to the popular isotropic Matérn family. One realization of this process is observed at n sites $\mathbf{s}_i, i = 1, \cdots, n$, which are a **subset** of the $n_1 \times n_2$ regularly spaced locations $\bar{\mathbf{s}}_k$ in $[0,1] \times [0,1]$, with an additive Gaussian white noise whose variance is σ_0^2 (this noise can model either suspected homoscedastic measurement errors or a nugget effect added to Z, see e.g. Zhang and Zimmerman (2007) and references therein). Using a standard lexicographic ordering, the observations thus form a vector $\boldsymbol{y}$ of size n whose law is Gaussian :

$$\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \tau_0^2 \boldsymbol{R}(\theta_0) + \sigma_0^2 \boldsymbol{I}_{n,n}) \text{ where } \boldsymbol{R}(\theta) = \boldsymbol{J}_{n,n_1\times n_2} \tilde{\boldsymbol{R}}(\theta) \boldsymbol{J}_{n,n_1\times n_2}^T$$

with $\boldsymbol{J}_{n,n_1\times n_2}$ denoting the rectangular, "with coefficients $\in \{0,1\}$", incidence matrix, $\boldsymbol{I}_n$ the identity matrix, and $\tilde{\boldsymbol{R}}(\theta)$ the autocorrelation matrix of the gridded process i.e. the block Toeplitz matrix (with n_1^2 Toeplitz square blocks, each of size $n_2 \times n_2$) whose coefficients are given by

$$\left[\tilde{\boldsymbol{R}}(\theta)\right]_{j,k} := \rho_{\nu,\theta}(||\bar{\mathbf{s}}_j - \bar{\mathbf{s}}_k||), \ j,k = 1, \cdots, n_1 \times n_2,$$

$||\cdot||$ being the Euclidean norm and $\rho_{\nu,\theta}$ the Matérn function

$$\rho_{\nu,\theta}(x) = \frac{(\theta x)^\nu}{\Gamma(\nu) 2^{\nu-1}} K_\nu(\theta x), \quad x > 0, \ \theta > 0,$$

where K_ν is the modified Bessel function of the second kind of order $\nu > 0$.

We here use the parameters b, θ, σ, with $b_0 = \frac{\tau_0^2}{\sigma_0^2}$ called the true signal-to-noise (SNR). The **microergodic-parameter** is $c = b\sigma^2\theta^{2\nu}$.

In all the alternatives to ML we consider, the signal-variance τ_0^2 is "estimated", **given a candidate value for the noise-variance** σ^2, by the naive "bias corrected" **empirical variance** :

$\hat{\tau}^2_{\mathrm{EV}|\sigma} := n^{-1}\boldsymbol{y}^T\boldsymbol{y} - \sigma^2$ (and the "SNR estimate" is $\hat{b}_{\mathrm{EV}|\sigma} := \frac{\hat{\tau}^2_{\mathrm{EV}|\sigma}}{\sigma^2}$).

Proposed estimators VEZ and LE$_3$

When the true $\sigma_0{}^2$ is known, CGEM-EV$|\sigma_0$ (which stands for "Conditional Gibbs-Energy Mean and Empirical Variance") is quite efficient, both statistically and computationally (using a fast randomized-trace), for many common problems (see [5,6]). Recall that CGEM-EV$|\sigma$ consists of solving (a fixed-point algorithm is used in [2,3,4,7]) the equation in θ defined by

$$\mathrm{LE}_1\left(\theta \left| \hat{b}_{\mathrm{EV}|\sigma}, \sigma\right.\right) = 0 \quad \text{with} \quad \mathrm{LE}_1(\theta|b,\sigma) := \frac{1}{n}\left\langle \boldsymbol{A}_{b,\theta}\,\mathbf{y}, \boldsymbol{R}(\theta)^{-1}\boldsymbol{A}_{b,\theta}\,\mathbf{y}\right\rangle - b\,\sigma^2\,\frac{1}{n}\,\mathrm{tr}(\boldsymbol{A}_{b,\theta}) \qquad (1),$$

where the "smoothing matrix" $\boldsymbol{A}_{b,\theta} := b\,\boldsymbol{R}(\theta)\,(b\,\boldsymbol{R}(\theta) + \boldsymbol{I})^{-1}$ (a.k.a. a Kalman smoother); let us denote "the" root of (1) with $\sigma = \sigma_0$ (the smallest one in case of multiple roots) by $\hat{\theta}_{\mathrm{CGEMEV}|\sigma_0}$.

- The first extension of CGEM-EV to the case of unknown σ_0 is a simple (and fast!) "plugin" method: one computes a nonparametric estimator of σ^2, say $\hat{\sigma}^2_{\mathrm{VEZ}}$, by **extrapolating at 0_+ the (classic or robust) semi-variogram function as a function of the lag** (proposed by [9]) and substitutes $\hat{\sigma}^2_{\mathrm{VEZ}}$ for $\sigma_0{}^2$ in the CGEM-EV$|\sigma_0$ equation in θ, producing $\hat{\theta}_{\mathrm{CGEMEV}|\hat{\sigma}_{\mathrm{VEZ}}}$.

- The second possible extension consists of **adding to (1) a well-known equation satisfied by the ML estimate of σ^2** when the signal-to-noise ratio b and θ are fixed; precisely, σ is the root of

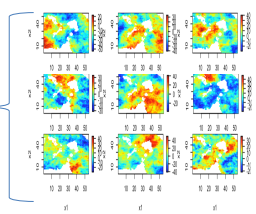
$$\mathrm{LE}_3(\sigma|b,\theta) := \frac{1}{n}\left\langle \mathbf{y}, (\boldsymbol{I} - \boldsymbol{A}_{b,\theta})\,\mathbf{y}\right\rangle - \sigma^2 \quad (2).$$

N.B. : In fact **LE$_3(\theta|b,\sigma)$** (resp. LE$_1(\sigma|b,\theta)$) **is proportional to the partial derivative of the log-likelihood w.r.t. the 3rd** (resp. 1st) **parameter**.

A classic bisection-search of σ, can be used to solve the concentrated CGEM-EV-LE$_3$ equation $\mathrm{LE}_3\left(\sigma\left|\hat{b}_{\mathrm{EV}|\sigma}, \hat{\theta}_{\mathrm{CGEMEV}|\sigma}\right.\right) = 0$ in σ, where an inner iteration is used to compute $\hat{\theta}_{\mathrm{CGEMEV}|\sigma}$. The final $\hat{\theta}$-iterate, thus root of (1) and (2), will be denoted by $\hat{\theta}_{\mathrm{CGEMEVLE}_3}$.

Experiments setting, VEZ-tuning

- $\nu = 1/2$
- $n_1 = n_2 = 54$. about 20% missing data (fixed locations) : n=2300
- signal-to-noise ratios b_0 $\left(= \frac{\mathrm{var}(Z)}{\mathrm{var}(\epsilon)}\right)$ chosen among: $7^2, 10^2, 15^2, 33^2$
- For each b_0 , inverse-range parameter $\theta_0 = 1, 3, 6, 12, 24, 48$

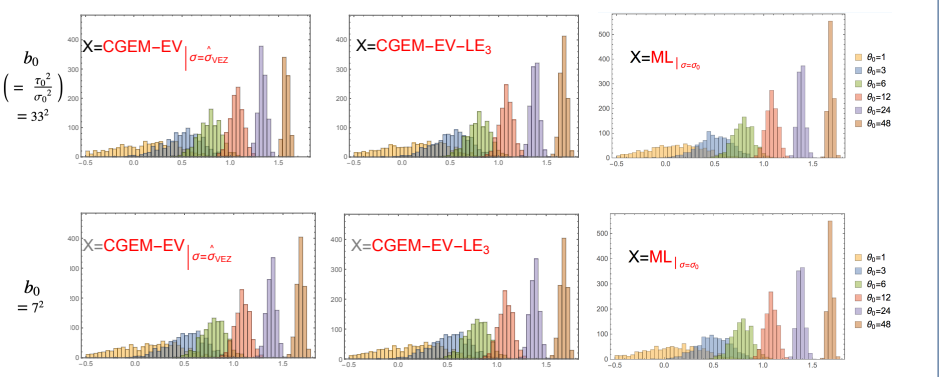


6 replicates for $\theta_0 = 3$, $b_0 = 15^2$

We use the "best performing" VEZ pour this setting, precisely:
- classic variogram,
- maximum-lag := 3/54

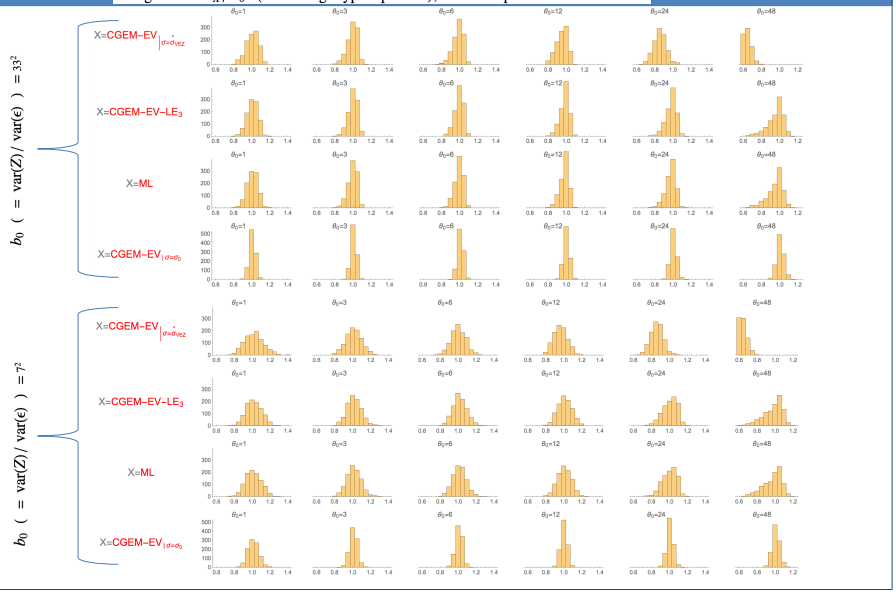
Estimation of the inverse-range parameter θ_0 .

histograms of $\log_{10}(\hat{\theta}_X)$ (1000 image-type replicates), observed pixels number n=2300



Estimation of the micro-ergodic parameter c_0

histograms of $\hat{c}_X / c_0$ (1000 image-type replicates), observed pixels number n=2300



Conclusions

- Concerning the estimation of the range-parameter θ_0, estimating the nugget-variance via CGEM-EV-LE$_3$ only causes a very small additional inaccuracy, even when compared with "ML given the true σ_0". The simplest CGEMEV $|\hat{\sigma}_{\mathrm{VEZ}}$ is almost as good, except for the largest θ_0's.

- Concerning the estimation of the micro-ergodic-parameter $c_0 = b_0\,\sigma_0^2\,\theta_0^{2\nu}$, CGEMEV $|\hat{\sigma}_{\mathrm{VEZ}}$ is clearly not efficient for the two largest θ_0's (and also for $\theta_0 \geq 6$, in case of strong SNR). But CGEM-EV-LE$_3$ is always nearly as efficient as ML (also note that the ideal "ML given the true σ_0" would much better estimate c_0) .

The R-package CGEMEV can efficiently compute $\hat{\theta}_{\mathrm{CGEMEV}|\sigma}$ for much larger image-type data (any product $\boldsymbol{A}_{b,\theta}\,\mathbf{x}$ using a fast, matrix free, preconditioned iterative algorithm, like FSAI-PCG) : → ongoing work is an assessment [1] of various versions of the "bisection, with inner iterations" mentioned above for computing $\hat{\theta}_{\mathrm{CGEMEVLE}}$.

this poster

github.com for R-package

References

[1] Drouilhet R., Girard D.A. (2019) "Simple and efficient procedures for fitting isotropic Matérn covariance models to large (noisy) data sets " To appear

[2] Girard D.A. (2014) "Estimating a Centered Ornstein-Uhlenbeck Process under Measurement Errors," *Wolfram Demonstrations Project*, Published: July 1, 2014.

[3] Girard D.A. (2015) "Three Alternatives to the Likelihood Maximization for Estimating a Centered Matérn (3/2) Process, " *Wolfram Demonstrations Project*,

[4] Girard D.A. (2015) " Estimating a Centered Matérn (1) Process: Three Alternatives to the Likelihood Maximization via conjugate gradient linear solvers" *Wolfram Demonstrations Project*,

[5] Girard D.A. (2016 & 2019). "Asymptotic Near-Efficiency of the 'Gibbs-Energy and Empirical-Variance' Estimating Functions for Fitting Matérn Models, I: Densely sampled processes, & II: Accounting for measurement errors via conditional GE mean," *Stat. Prob. Letters* & arxiv.org/pdf/0909.1046v3.pdf.

[6] Girard D.A. (2017) . "Efficiently estimating some common geostatistical models by 'energy-variance matching' or its randomized 'conditional-mean' versions," *Spatial Statistics*, 21(Part A), 1-26

[7] Girard D.A. (2018) "Estimators of a Noisy Centered Ornstein-Uhlenbeck Process and Its Noise Variance, " *Wolfram Demonstrations Project*,

[8] Kaufman C.G., Shaby, B.A. (2013). "The Role of the Range Parameter for Estimation and Prediction in Geostatistics," *Biometrika* 100 (2), pp. 473-484.

[9] Katzfuss and N. Cressie, "Bayesian Hierarchical Spatio-temporal Smoothing for Very Large Datasets," *Environmetrics* , 23(1), 2012 pp. 94-107.

[10] Zhang H. (2012). "Asymptotics and Computation for Spatial Statistics," in *Advances and Challenges in Space-time Modelling of Natural Events (Lecture Notes in Statistics, Vol. 207)* (E. Porcu, J. M. Montero, and M. Schlather, eds.), New York: Springer, pp. 239-252.

[11] Zhang, H., Zimmerman, D.L., 2007. "Hybrid estimation of semivariogram parameters," *Mathematical Geology* 39 (2), 247-260.

Work developed in the framework of Grenoble Alpes Data Institute, supported by the French National Research Agency under the " Investissements d'avenir" program (ANR-15-IDEX-02).